Appendix 32. Model Archival Summary for Enterococci Bacteria Density at U.S. Geological Survey Site 06892350, Kansas River at De Soto, Kansas, during September 2013 through September 2019

This model archival summary summarizes the enterococci bacteria (ENT; U.S. Geological Survey [USGS] parameter code 90909) density model developed to compute 15-minute ENT densities from September 2013 onward. This model supersedes all previous models.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Site and Model Information

Site number: 06892350
Site name: Kansas River at De Soto, Kansas
Location: Lat 38°59'00", long 94°57'52" referenced to North American Datum of 1927, in NE 1/4 SE 1/4 SE 1/4 sec.28, T.12 S., R.22 E., Leavenworth County, Kans., hydrologic unit 10270104.

Equipment: A YSI 6600 water-quality monitor equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, and turbidity (TBY) was installed from August 2012 through June 2014. A Xylem YSI EXO2 water-quality monitor equipped with sensors for water temperature, specific conductance, dissolved oxygen, pH, TBY, and chlorophyll and phycocyanin fluorescence was installed during June 2014 through September 2019. A Hach Nitratax plus sc sensor (5-millimeter path length) that monitors ultraviolet (UV) nitrate concentrations was installed from June 2013 through September 2019. The monitors were housed in side-by-side 4-inch-diameter galvanized steel pipes. Readings from the water-quality and nitrate plus nitrite monitors were recorded every 15 minutes and transmitted by way of satellite, hourly.

Date model was created: April 14, 2020

Model calibration data period: September 23, 2013, through September 24, 2019

Model application date: September 23, 2013, onward

Model-Calibration Dataset

All data were collected using USGS protocols (Wagner and others, 2006; U.S. Geological Survey, variously dated) and are stored in the National Water Information System (U.S. Geological Survey, 2020) database and available to the public. Ordinary least squares analysis was used to develop regression models using R programming language (R Core Team, 2020). Potential explanatory variables that were evaluated individually and in combination included streamflow, water temperature, specific conductance, dissolved oxygen, pH, TBY, chlorophyll and phycocyanin fluorescence, and UV nitrate sensor data. The maximum time span between two continuous data points used for interpolation was 2 hours (in order to preserve the sample dataset, field monitor averages obtained during sample collection were used for model development data if no continuous data were available or if gaps larger than 1 hour in the continuous data record resulted in missing interpolated data). Seasonal components (sine and cosine variables) were also evaluated as potential explanatory variables.

The final selected regression model was based on 76 concurrent measurements of ENT density and sensor-measured TBY during September 23, 2013, through September 24, 2019. Samples were collected throughout the range of continuously observed hydrologic conditions. No samples had densities below laboratory detection limits. Thirty sample densities were qualified as "estimated." One sample, from December 14, 2015, exceeded the laboratory detection limit (greater than 20,000 colony forming units) and was removed from the model calibration dataset due to a high level of uncertainty and high leverage it would have on the final model. Summary statistics and the complete model-calibration dataset are provided below. Potential outliers were identified using the methods described in Rasmussen and others (2009). Additionally, studentized residuals from the final model were inspected for values greater than three or less than negative three. Values outside of that range were considered potential outliers and were investigated. All potential outliers were not found to have errors associated with collection, processing, or analysis and were therefore considered valid.

This model is specific to the Kansas River at De Soto, Kans., during this study period and cannot be applied to data collected from other sites on the Kansas River or data collected from other waterbodies.

Enterococci Bacteria Sampling Details

Indicator bacteria samples typically were collected either from the downstream side of the bridge or instream within 100 feet of the bridge. The grab sample collection method with weighted basket was used for all indicator bacteria samples (contrary to the equal-

width-increment collection method used for all other analytes; U.S. Geological Survey, variously dated). During July 2012 through June 2017, grab samples were collected every 2 weeks during March through October, once a month during November through February, and during selected reservoir release and runoff events. During July 2017 through September 2019, grab samples were collected on a monthly to bimonthly basis, depending on flow conditions. An open-mouth bottle with weighted-basket sampler was used. Additional detail on sample collection is available in Foster and Graham (2016) and Graham and others (2018). Samples were analyzed for ENT density at the USGS Kansas Water Science Center in Lawrence, Kans.

Model Development

Ordinary least squares regression analysis was done using R programming language (R Core Team, 2020) to relate discretely collected ENT density to sensor-measured TBY. The distribution of residuals was examined for normality, and the plots of residuals (the difference between the measured and computed values) were examined for homoscedasticity (departures from zero did not change substantially over the range of computed values). Previously published explanatory variables were also strongly considered for continuity.

TBY was selected as a good surrogate for ENT based on residual plots, coefficient of determination (R^2), and model standard percentage error. Values for all the aforementioned statistics were computed and are included below along with all relevant sample data and additional statistical information.

Model Summary

The following is a summary of final regression analysis for ENT density at USGS site 06892350:

ENT density-based model:

$$\log ENT = 1.38 \times \log TBY - 0.295$$

where

log = logarithm base 10;

ENT = enterococci bacteria density, in colonies per 100 milliliters; and

TBY = turbidity, in formazin nephelometric units.

TBY makes physical and statistical sense as an explanatory variable for ENT because of its positive correlation with suspended material to which fecal indicator bacteria can physically bind.

The logarithmically (log) transformed model may be retransformed to the original units so that ENT can be calculated directly. The retransformation introduces a bias in the calculated constituent. This bias may be corrected using Duan's bias correction factor (BCF; Duan, 1983). For this model, the calculated BCF is 1.62. The retransformed model, accounting for BCF is as follows:

$$ENT = 1.62 \times (TBY^{1.38} \times 10^{-0.295})$$

Previous Models

Start Year	End Year	Model Equation	Reference
2012	2019	logENT = 1.39logTBY + 0.211sin(2πD) + 0.214cos(2πD) - 0.292	Foster and Graham (2016)
1999	2003	logENT = 1.64logTBY – 0.768	Rasmussen and others (2005)

Model Statistics, Data, and Plots

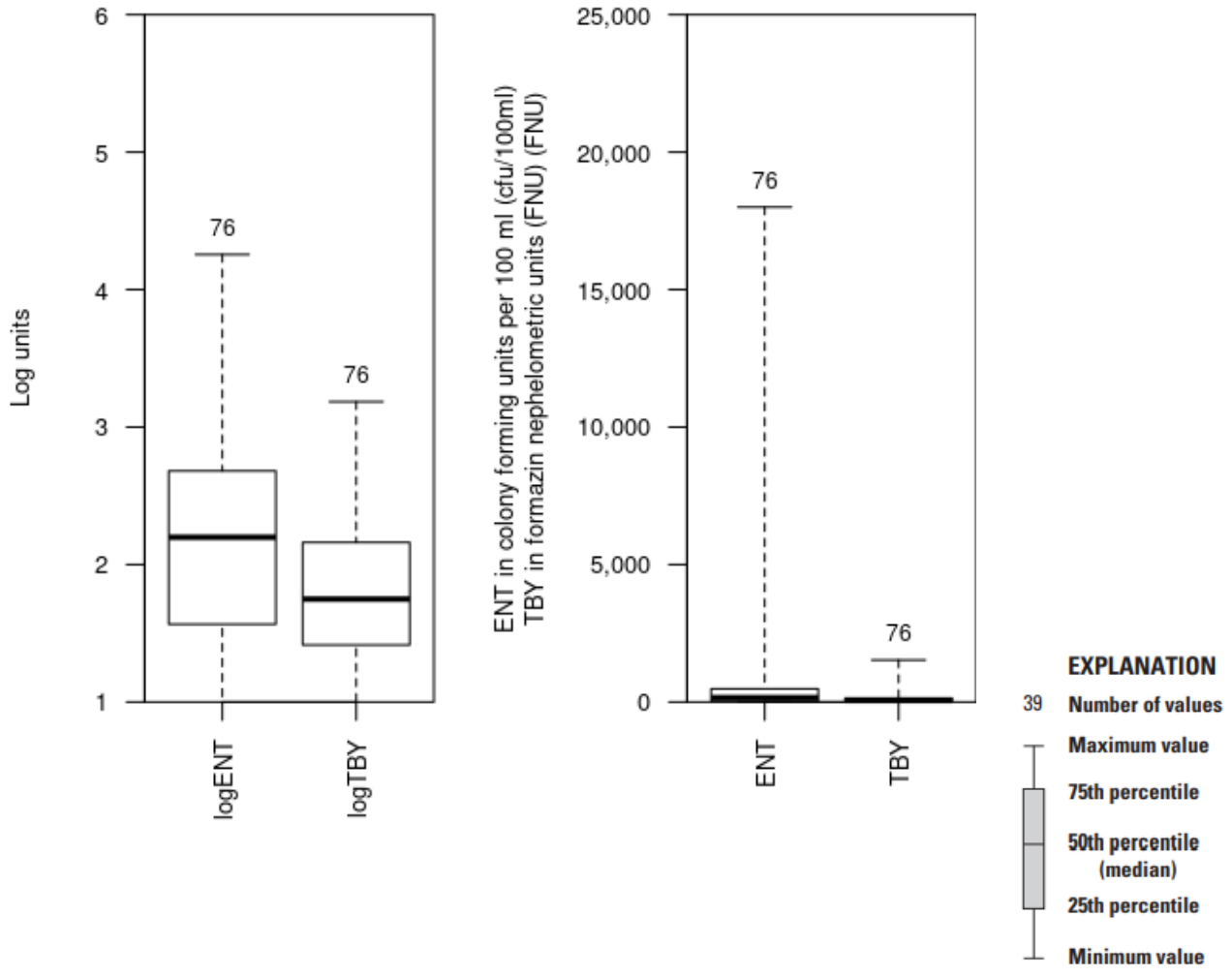
Model

logENT = + 1.38 * logTBY - 0.295

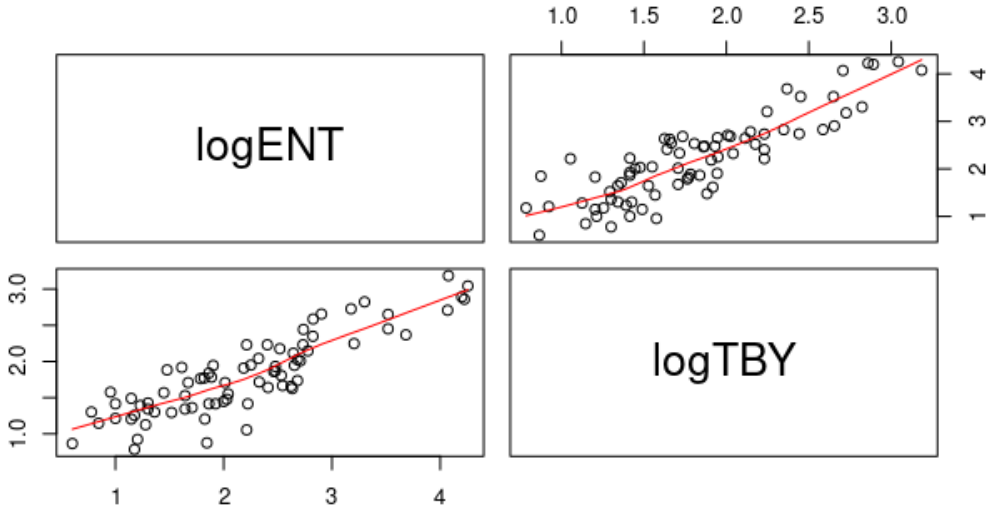
Variable Summary Statistics

	logENT	ENT	logTBY	TBY
Minimum	0.602	4	0.785	6.1
1st Quartile	1.570	37	1.410	26.0
Median	2.200	158	1.750	56.2
Mean	2.200	1350	1.810	153.0
3rd Quartile	2.680	480	2.160	145.0
Maximum	4.260	18000	3.180	1530.0

Box Plots



Exploratory Plots



Red line shows the locally weighted scatterplot smoothing (LOWESS).

The x- and y-axis labels for a given bivariate plot are defined by the intersecting row and column labels.

Basic Model Statistics

Number of Observations	76
Standard error (RMSE)	0.433
Average Model standard percentage error (MSPE)	117
Coefficient of determination (R^2)	0.757
Adjusted Coefficient of Determination (Adj. R^2)	0.754
Bias Correction Factor (BCF)	1.62

Explanatory Variables

	Coefficients	Standard Error	t value	Pr(>|t|)
(Intercept)	-0.295	0.1720	-1.72	9.03e-02
logTBY	1.380	0.0908	15.20	1.97e-24

Correlation Matrix

	Intercept	E.vars
Intercept	1.000	-0.957
E.vars	-0.957	1.000

Outlier Test Criteria

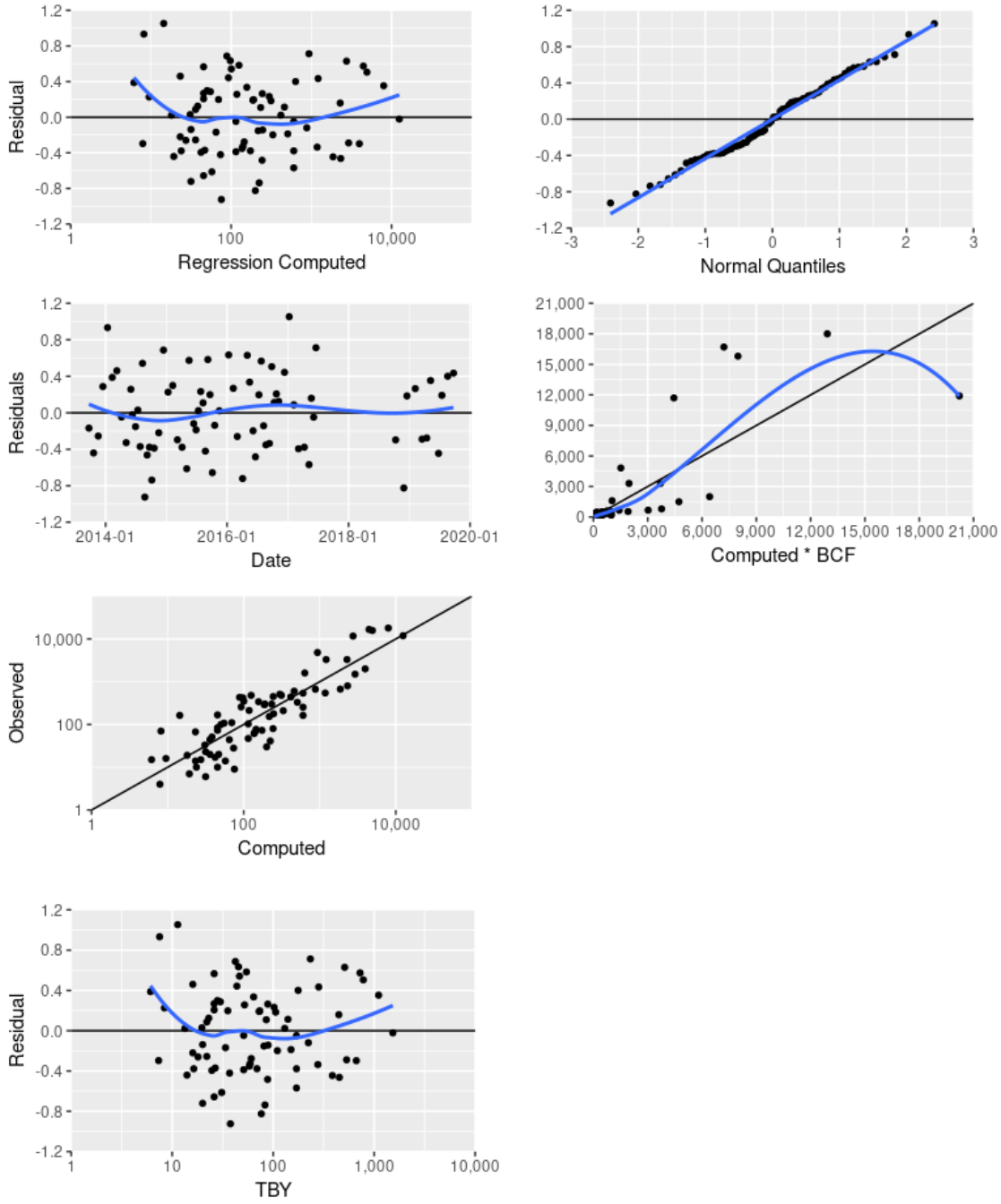
Leverage	Cook's D	DFFITS
0.0789	0.1943	0.3244

Flagged Observations

	logENT	Estimate	Residual	Standard Residual	Studentized Residual	Leverage	Cook's D	DFFITS
201401130800	1.85	0.912	0.933	2.2100	2.2800	0.0516	0.13300	0.5310
201406111600	4.08	4.100	-0.021	-0.0512	-0.0508	0.0965	0.00014	-0.0166
201505181530	4.22	3.650	0.574	1.3700	1.3800	0.0618	0.06170	0.3540
201605020930	4.07	3.440	0.629	1.4900	1.5000	0.0487	0.05690	0.3400

201701090950	2.21	1.160	1.050	2.4800	2.5700	0.0383	0.12300	0.5140
201905091110	4.26	3.900	0.353	0.8510	0.8500	0.0803	0.03160	0.2510

Statistical Plots



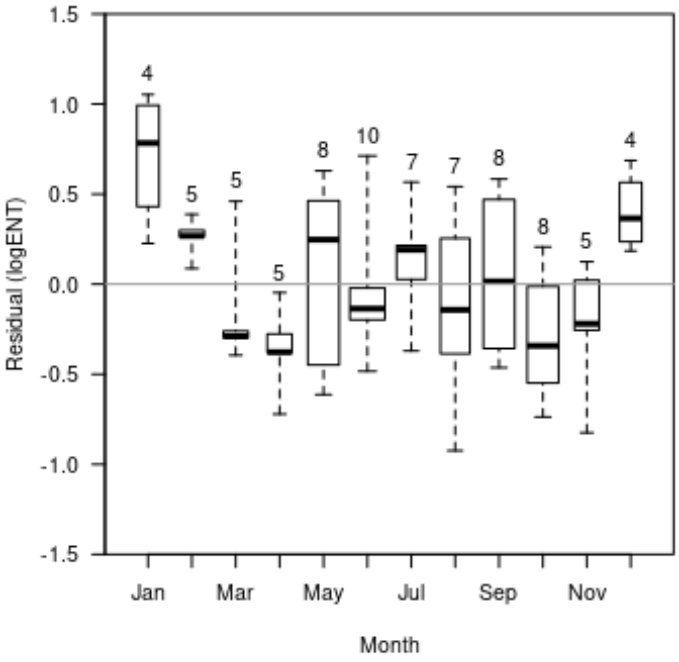
First row (left): Residual ENT related to regression computed ENT with local polynomial regression fitting, or locally estimated scatterplot smoothing (LOESS), indicated by the blue line.

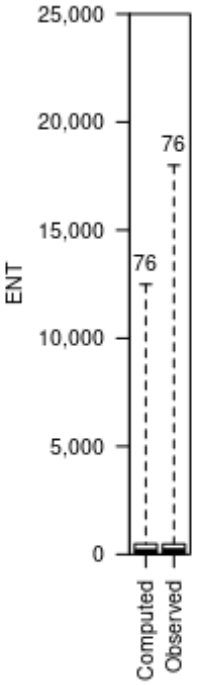
First row (right): Residual ENT related to the corresponding normal quantile of the residual with simple linear regression, indicated by the blue line.

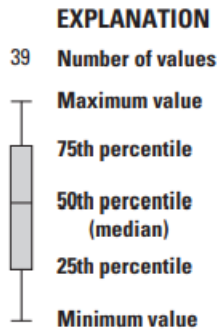
Second row: Residual ENT related to date (left) and regression computed ENT multiplied by the BCF (right) with LOESS, indicated by the blue line.

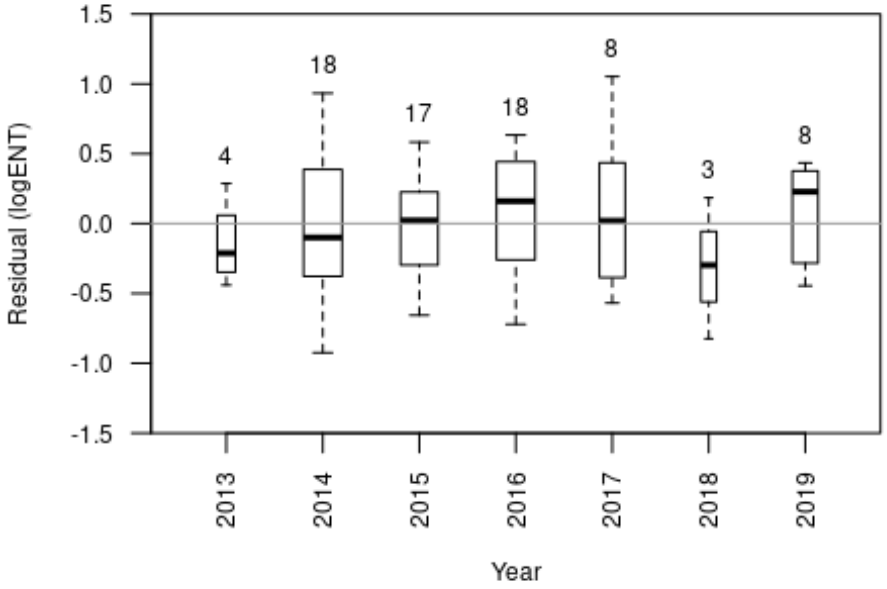
Third row: Observed ENT related to regression computed ENT.

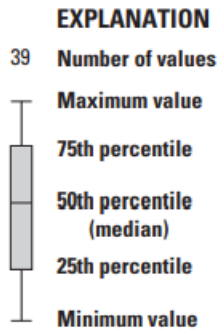
Fourth row: Residual ENT related to TBY with LOESS, indicated by the blue line.


Residual (logENT)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
4
5
5
5
8
10
7
7
8
8
5
4
Jan
Mar
May
Jul
Sep
Nov
Month

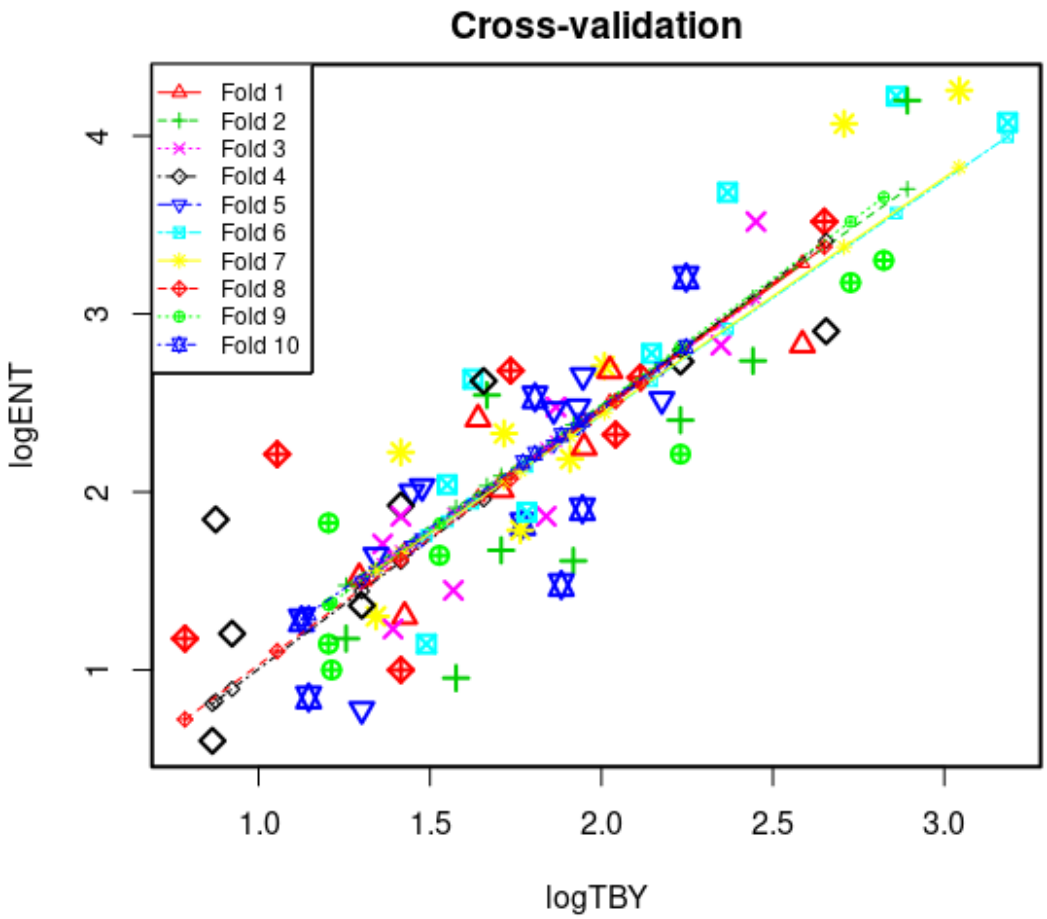

ENT
25,000
20,000
15,000
10,000
5,000
0
76
76
Computed
Observed


EXPLANATION
39
Number of values
Maximum value
75th percentile
50th percentile
(median)
25th percentile
Minimum value


Residual (logENT)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
4
18
17
18
8
3
8
2013
2014
2015
2016
2017
2018
2019
Year


EXPLANATION
39
Number of values
Maximum value
75th percentile
50th percentile
(median)
25th percentile
Minimum value

Cross-Validation



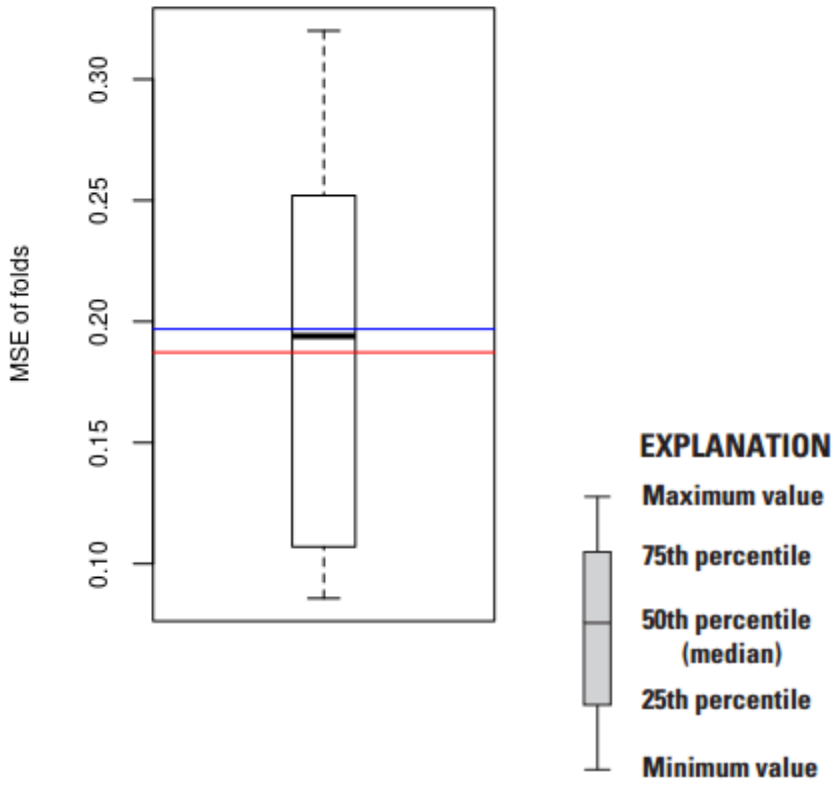

```
Fold - equal partition of the data (10 percent of the data).

Large symbols – observed value of a data point removed in a fold.

Small symbols – recomputed value of a data point removed in a fold.

Recomputed regression lines – adjusted regression line with one fold removed.


              Minimum MSE of folds:  0.0857
                 Mean MSE of folds:  0.1970
               Median MSE of folds:  0.1940
              Maximum MSE of folds:  0.3200
 (Mean MSE of folds) / (Model MSE):  1.0500
```



Red line - Model MSE

Blue line - Mean MSE of folds

Model-Calibration Dataset

	Date	logENT	logTBY	ENT	TBY	Computed logENT	Computed ENT	Residual	Normal Quantiles	Censored Values
1	2013-09-23	1.64	1.53	44	33.7	1.81	105	-0.168	-0.249	--
2	2013-10-21	0.845	1.15	7	14	1.29	31.2	-0.441	-1.08	--
3	2013-11-18	1.3	1.34	20	22	1.56	58.3	-0.255	-0.388	--
4	2013-12-16	2.03	1.48	107	30	1.74	89.4	0.287	0.652	--
5	2014-01-13	1.85	0.875	70	7.5	0.912	13.2	0.933	2.03	--
6	2014-02-10	1.18	0.785	15	6.1	0.788	9.94	0.388	0.825	--
7	2014-03-10	1.83	1.2	67	16	1.37	37.6	0.46	1.03	--
8	2014-04-07	2.01	1.71	103	51	2.06	186	-0.047	-0.0494	--
9	2014-05-05	1.82	1.77	66	59	2.15	227	-0.328	-0.612	--
10	2014-06-02	2.33	1.72	213	52	2.07	191	0.257	0.534	--
11	2014-06-11	4.08	3.18	11900	1530	4.1	20200	-0.021	-0.0164	--
12	2014-06-30	2.18	1.91	153	81	2.34	352	-0.152	-0.215	--
13	2014-07-14	1.52	1.29	33	19.7	1.49	49.9	0.0293	0.0823	--
14	2014-07-28	1.3	1.43	20	26.7	1.67	76	-0.371	-0.736	--
15	2014-08-11	2.54	1.67	350	46.3	2	163	0.542	1.15	--
16	2014-08-25	0.954	1.58	9	37.7	1.88	122	-0.924	-2.41	--
17	2014-09-08	2.9	2.66	800	452	3.37	3760	-0.463	-1.21	--
18	2014-09-22	1.86	1.84	73	69	2.24	282	-0.378	-0.779	--
19	2014-10-06	1.61	1.92	41	82.8	2.35	362	-0.737	-1.82	--
20	2014-10-20	1.67	1.71	47	51	2.06	186	-0.388	-0.921	--

21	2014-11-17	1.15	1.2	14	16	1.37	37.6	-0.22	-0.353	--
22	2014-12-15	2.63	1.63	430	42.2	1.95	143	0.687	1.67	--
23	2015-01-12	1.2	0.923	16	8.37	0.978	15.4	0.226	0.46	--
24	2015-02-09	2	1.45	100	28	1.7	81.3	0.299	0.693	--
25	2015-03-09	0.602	0.865	4	7.33	0.898	12.8	-0.296	-0.534	--
26	2015-04-06	1	1.21	10	16.3	1.38	38.6	-0.378	-0.872	--
27	2015-05-04	1.15	1.49	14	30.9	1.76	93	-0.613	-1.45	--
28	2015-05-18	4.22	2.86	16700	724	3.65	7210	0.574	1.28	--
29	2015-06-15	2.82	2.35	667	223	2.94	1420	-0.119	-0.115	--
30	2015-06-29	2.52	2.18	330	150	2.71	822	-0.187	-0.283	--
31	2015-07-13	2.64	2.11	440	130	2.62	675	0.0233	0.0494	--
32	2015-07-27	2.71	2.01	510	102	2.47	483	0.233	0.497	--
33	2015-08-10	2.48	1.93	300	85.3	2.37	378	0.109	0.149	--
34	2015-08-24	1.45	1.57	28	37	1.87	119	-0.421	-1.03	--
35	2015-09-08	2.68	1.74	480	54.3	2.1	203	0.583	1.36	--
36	2015-09-21	2.04	1.55	110	35.5	1.84	113	0.198	0.388	--
37	2015-10-05	1	1.41	10	26	1.66	73.4	-0.656	-1.55	--
38	2015-10-19	1.36	1.3	23	20	1.5	51.1	-0.138	-0.149	--
39	2015-11-16	1.28	1.12	19	13.3	1.26	29.2	0.0223	0.0164	--
40	2016-01-11	2.62	1.66	420	45.3	1.99	158	0.634	1.55	--
41	2016-02-08	1.92	1.41	84	26	1.66	73.4	0.268	0.612	--
42	2016-03-03	1.18	1.26	15	18	1.44	44.2	-0.26	-0.424	--
43	2016-04-04	0.778	1.3	6	20	1.5	51.1	-0.721	-1.67	--
44	2016-05-02	4.07	2.71	11700	510	3.44	4440	0.629	1.45	--
45	2016-05-16	2.53	1.81	340	64	2.2	254	0.336	0.736	--
46	2016-06-06	2.32	2.04	210	110	2.52	536	-0.198	-0.318	--
47	2016-06-20	1.9	1.94	80	88	2.39	394	-0.483	-1.28	--
48	2016-07-11	2.48	1.87	300	73.7	2.28	308	0.197	0.353	--
49	2016-07-25	2.22	1.41	167	26	1.66	73.4	0.566	1.21	--
50	2016-08-08	2.25	1.95	178	89	2.39	400	-0.143	-0.182	--
51	2016-08-22	1.79	1.76	61	58	2.14	222	-0.352	-0.693	--
52	2016-09-12	2.74	2.44	545	277	3.07	1910	-0.336	-0.652	--
53	2016-09-26	4.2	2.89	15800	780	3.69	7980	0.506	1.08	--
54	2016-10-11	2.78	2.15	600	140	2.66	748	0.114	0.182	--
55	2016-10-24	1.86	1.41	73	26	1.66	73.4	0.207	0.424	--
56	2016-11-07	1.71	1.36	51	23	1.58	62	0.125	0.215	--
57	2016-12-12	2.41	1.64	257	43.7	1.97	150	0.443	0.973	--
58	2017-01-09	2.21	1.05	163	11.3	1.16	23.3	1.05	2.41	--
59	2017-02-06	1.64	1.34	44	22	1.56	58.3	0.0871	0.115	--
60	2017-03-06	1.23	1.39	17	24.7	1.62	68.2	-0.394	-0.973	--
61	2017-04-10	2.4	2.23	253	170	2.78	977	-0.378	-0.825	--
62	2017-05-08	2.21	2.23	163	170	2.78	977	-0.569	-1.36	--
63	2017-05-22	3.52	2.65	3300	447	3.36	3700	0.159	0.249	--
64	2017-06-05	2.73	2.23	540	170	2.78	977	-0.0484	-0.0823	--
65	2017-06-19	3.68	2.37	4820	233	2.97	1510	0.713	1.82	--
66	2018-10-11	3.3	2.82	2000	666	3.6	6420	-0.297	-0.573	--
67	2018-11-29	1.48	1.88	30	76.4	2.3	324	-0.824	-2.03	--
68	2018-12-18	2.68	2.02	480	106	2.5	508	0.185	0.283	--
69	2019-02-06	2.65	1.95	450	88.3	2.39	396	0.264	0.573	--
70	2019-03-19	3.18	2.73	1500	533	3.47	4720	-0.289	-0.497	--
71	2019-04-16	1.89	1.78	77	60.6	2.16	236	-0.277	-0.46	--
72	2019-05-09	4.26	3.04	18000	1110	3.9	12900	0.353	0.779	--
73	2019-06-26	2.83	2.59	670	385	3.27	3020	-0.445	-1.15	--
74	2019-07-16	2.46	1.86	290	72.6	2.27	302	0.191	0.318	--

75	2019-08-20	3.2	2.25	1600	177	2.8	1030	0.4	0.872	--
76	2019-09-24	3.52	2.45	3300	282	3.08	1970	0.434	0.921	--

Definitions

Cook's D: Cook's distance (Helsel and others, 2020).

DFFITS: Difference in fits statistic (Helsel and others, 2020).

E.vars: Explanatory variables.

ENT: Enterococci, in colonies per 100 milliliters (90909).

Leverage: An outlier's measure in the x direction (Helsel and others, 2020).

LOESS: Local polynomial regression fitting, or locally estimated scatterplot smoothing (Helsel and others, 2020).

LOWESS: Locally weighted scatterplot smoothing (Cleveland, 1979; Helsel and others, 2020).

MSE: Model standard error (Helsel and others, 2020).

MSPE: Model standard percentage error (Helsel and others, 2020).

Probability(>|t|): The probability that the independent variable has no effect on the dependent variable (Helsel and others, 2020).

RMSE: Root mean square error (Helsel and others, 2020).

t value: Student's t value; the coefficient divided by its associated standard error (Helsel and others, 2020).

TBY: Turbidity, in formazin nephelometric units (63680).

References Cited

Cleveland, W.S., 1979, Robust locally weighted regression and smoothing scatterplots: Journal of the American Statistical Association, v. 74, no. 368, p. 829-836.

Duan, N., 1983, Smearing estimate—A nonparametric retransformation method: Journal of the American Statistical Association, v. 78, no. 383, p. 605–610. [Also available at https://doi.org/10.1080/01621459.1983.10478017.]

Foster, G.M., and Graham, J.L., 2016, Logistic and linear regression model documentation for statistical relations between continuous real-time and discrete water-quality constituents in the Kansas River, Kansas, July 2012 through June 2015: U.S. Geological Survey Open-File Report 2016–1040, 27 p., accessed July 2020 at https://doi.org/10.3133/ofr20161040.

Graham, J.L., Foster, G.M., Williams, T.J., Mahoney, M.D., May, M.R., and Loftin, K.A., 2018, Water-quality conditions with an emphasis on cyanobacteria and associated toxins and taste-and-odor compounds in the Kansas River, Kansas, July 2012 through September 2016: U.S. Geological Survey Scientific Investigations Report 2018–5089, 55 p. [Also available at https://doi.org/10.3133/sir20185089.]

Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., and Gilroy, E.J., 2020, Statistical methods in water resources: U.S. Geological Survey Techniques and Methods, book 4, chap. A3, 458 p. [Also available at https://doi.org/10.3133/tm4a3.] [Supersedes USGS Techniques of Water-Resources Investigations, book 4, chap. A3, ver. 1.1.]

R Core Team, 2020, R—A language and environment for statistical computing, version 4.0.3: Vienna, Austria, R Foundation for Statistical Computing, accessed December 2020 at https://www.R-project.org/.

Rasmussen, P.P., Gray, J.R., Glysson, G.D., and Ziegler, A.C., 2009, Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity sensor and streamflow data: U.S. Geological Survey Techniques and Methods, book 3, chap. C4, 53 p. [Also available at https://doi.org/10.3133/tm3C4.]

Rasmussen, T.J., Ziegler, A.C., and Rasmussen, P.P., 2005, Estimation of constituent concentrations, densities, loads, and yields in lower Kansas River, northeast Kansas, using regression models and continuous water-quality monitoring, January 2000 through December 2003: U.S. Geological Survey Scientific Investigations Report 2005–5165, 117 p. [Also available at https://doi.org/10.3133/sir20055165.]

U.S. Geological Survey, 2020, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed April 2020 at https://doi.org/10.5066/F7P55KJN.

U.S. Geological Survey, variously dated, National field manual for the collection of water-quality data: U.S. Geological Survey Techniques of Water-Resources Investigations, book 9, chaps. A1–A9 [variously paged], accessed July 2020 at https://water.usgs.gov/owq/FieldManual/.

Wagner, R.J., Boulger, R.W., Jr., Oblinger, C.J., and Smith, B.A., 2006, Guidelines and standard procedures for continuous water-quality monitors—Station operation, record computation, and data reporting: U.S. Geological Survey Techniques and Methods, book 1, chap. D3, 51 p. plus 8 attachments. [Also available at https://doi.org/10.3133/tm1D3.] [Supersedes USGS Water-Resources Investigations Report 2000–4252.]